Reading Gray Literature as Texts. Semantic Mark-up of Museum Acquisition Catalogues

By Øyvind Eide and Jon Holmen, Unit for Digital Documentation, University of Oslo

Summary

Since the early 1990's, we have created SGML and XML encoded digital versions of printed acquisition catalogues at the Unit for Digital Documentation through projects such as the Museum Project. In this paper, we describe the various element groups we have defined as relevant in encoding such catalogues and how they can be used on texts in which the information is conveyed in a variety of ways. We show how an ODD document including the description and documentation is created, and demonstrate how this can be included in a formal description compatible with the TEI standard. We also discuss the import of texts encoded in this way into a CIDOC-CRM compatible database.

Introduction

The acquisition catalogues are important sources to historical information about the collections of Norwegian museums. Based on types of informations found to be important in reading such catalogues, we have developed several data descriptions in the form of SGML and eventually XML DTDs for the purpose of encoding acquisition catalogues (Holmen 1996), (Holmen forthcoming).

We are now in the process of implementing a data description for acquisition catalogue mark-up for which we will use the ODD system (TEI P5, chapter 27) developed in close connection to the Text Encoding Initiative (TEI web). Not only is this a practical way to create formal descriptions and documentation, helping us in understanding and conveying the information structures expressed in the texts, it also makes it easy to include the whole data description as a module to be used in TEI document descriptions.

Groups of elements

The paragraph

An acquisition catalogue contains descriptions of artefacts in a museum. The artefacts are grouped together in collections of varying size. Each of these collections is identified by an inventory number. Thus, the “inventory number” is the core unit of the catalogues. The expression “inventory number” is used for the whole description of the collection of artefacts as well as for the number identifying it. The inventory numbers are commonly divided into smaller parts identified by sub-numbers.

The section of a catalogue describing the artefacts being part of an inventory number is encoded with the **textPart** element. It contains a **mnr** element encoding the museum number followed by any number of **textPart** elements (for sub-numbers) and **artData** elements. This gives the following general structure:

```
<textPart> --> Information about one inventory number
  <mnr> --> The number itself
  <artData> --> Information about the artifact(s)
    ...
  <textPart> --> Information about one inventory sub-number
```

```
      <mnr> --> The sub-number itself
      <artData> --> Information about the artifact(s)
        ...
    <textPart> --> Information about one inventory sub-number
      <mnr> --> The sub-number itself
      <artData> --> Information about the artifact(s)
        ...
  <textPart> --> Information about one inventory number
    ...
```

The **artData** element encodes information related to an artefact or a group of artefacts. As the acquisition catalogues are running text, there are no strict order in which the various aspects of the artefacts are described. The encoding of the documents also have to be flexible. The following groups of information can be found in many different orders in encoded documents, so there are no strict order defined in the XML data description. On the other hand, some elements are only found in relation to certain other elements, giving some structure to the data description.

Objects

One group of elements relates directly to the objects in the museum as they are described in the catalogue. They are used to encode such aspects of the artefacts as numbers, form, decoration, material and measures. Examples of the use of such elements, along with the ones described in the following paragraphs, are included below.

Events

The second group of elements is used to encode events described directly in or entailed from the acquisition catalogue. The first group of event elements are used for events of transitions from one actor and place to another actor and place, noting that actors can be institution as well as human beings. The most general form is a general transfer. This can be a transfer of ownership, with yet more specialised form such as inheritance, gift and trade, but also deposition and unspecified income to the museum.

The second group of event elements are used to encode events with one actor taking place at one place. This includes the core elements of human life such as birth, marriage and death, as well as important aspects of the history of the artefact, such as actual production and use. The third group of event elements are used for internal events in the museum, i.e. conservation and exhibition.

The fourth group of event elements are used for non-factual events. These events are used for typical production or use that may or may not be actual. An example is the artefact «a collar used by protestant priests». The statement will be true even if the actual collar in the museum have never been used by any priest, as the description is a type description. Thus, these element are closely related to classification of artefacts.

Actors

The actor elements are used to encode actors described in the acquisition catalogues. First, we define the various types of actors to be encoded. These are persons, families, organisations, and groups in general.

To enrich this further, we encode attributes of the actors, such as age, sex, nationality and name. We also encode specific information such as title (king, madame, etc.) and social status (married, widow, etc.).

Other elements

Time is an important aspect in all museum activity. We have elements for historical time, such as period, as well as for museum time, used for acquisition, exhibitions and other museum events. We have also defined a chronology element used to define what happens before something else in situations where this is important to encode. This element uses the IDs of other elements to define the chronology.

Places are also encoded, and are commonly specified as being the place something is moved from or to in events with two places and actors.

An example

The following example shows an excerpt from a typical catalogue in which a painting, its production and the acquisition by the museum is described:

"Oil painted portrait in golden frame. Signed in right corner: "Both 1834" and must have been painted by Knud Andreassen Baade (1808-1879). Given to the collection by the inheritants after miss Abellone Gram at Midsø ved Stenkjer."

An XML encoded version of this text using the element types described above might look like the following:

```
<artData>
 <prodTech>Oil painted</prodTech>
 <artefact>
  <form>portrait</form>
 </artefact>
 in
 <prodTech>golden</prodTech>
 <form>frame</form>.
 <formDecor>Signed in right corner: "Both
  <fProd>
   <time>1834</time>
  </fProd>"
 </formDecor>
 and
 <fProd>must have been painted by
  <pers>
   <name>Knud Andreassen Baade</name> (
   <bYear>1808</bYear>-
   <dYear>1879</dYear>).
  </pers>
 </fProd>
 <gift>Given to the collection by
  <inheritance>
   <group>the inheritants after
    <person>
     <title>miss</title>
     <name>Abellone Gram</name>
    </person>
   </group>
  </inheritance>
  at
  <place>Midsø ved Stenkjer</place>
 </gift>
</artData>
```

Overlapping hierarchies

One consequence of the fact that we are encoding prose description is that the descriptions are not

standardised, they come in a variety of forms. This necessitates the loose data definitions described above. Another consequence is the fact that the information we encode often do not nest in a way that is needed for pure XML element encoding. This is a well known problem in text encoding and there are several ways to handle it (TEI 2005, chapter 31). We use element fragmenting and reconstruction of virtual elements by id/idref links to establish structures that are not expressed in the XML tree. An example of the use of such linking can be found in this example, showing a similar set of information as the former example, but structured differently:

"During examinations in the museum we found two paintings, painted by Knud Andreassen Baade and Henrik Soerensen. Both are painted at the same place, Hokksund, and have the same nature view, but Soerensen's is painted in 1934 and Baade's in 1834. Both oil paintings were given to the museum at the same place by the two painters' grandchildren. Baade's in 1910 and Soerensen's in 1978."

An encoding of this excerpt can not be based solely on the relations between elements being expressed by the tree structure of the XML version, but have to use explicit links expressed as attributes to show the relations between various elements:

```
<artData objid="obj1 obj2">
 During examinations in the museum we found two
  <artType>paintings</artType>, painted by
  <fprod objref="obj1" placeref="s1" timeref="t2">
    <pers id="p1">
      <name>Knud Andreassen Baade</name>
    </pers>
  </fprod>
 and
  <fprod objref="obj2" placeref="s1" timeref="t1">
    <pers id="p2">
      <name>Henrik Soerensen</name>
    </pers>
  </fprod>.
   Both are painted at the same place,
  <place id="s1">Hokksund</place>,
 and have the same nature view, but
  <pers id="p2">
    <name>Soerensen's</name>
  </pers>
 is painted in
  <time id="t1">1934</time>
 and
  <pers id="p1">
    <name>Baade's</name>
  </pers>
 in
  <time id="t2">1834</time>.
  <gift oref="obj1" timeref="t3" placeref="s1" fromid="p3"
toid="m1"></gift>
  <gift oref="obj2" timeref="t4" placeref="s1" fromid="p4"
toid="m1"></gift>
 Both oil paintings were given to
  <org id="m1">the museum</org>
 at the same place by
  <pers id="p3" gcref="p1"></pers>
  <pers id="p4" gcref="p2"></pers>
 the two painters' grandchildren.
  <pers id="p1">
    <name>Baade's</name>
  </pers>
 in
```

```
  <time id="t3">1910</time>
 and
  <pers id="p2">
    <name>Soerensen's</name>
  </pers>
 in
  <time id="t4">1978</time>.
</artData>
```

The general rule for interpretation of the elements is that all elements take as their relational elements the other elements to which they have idref links. In addition, they take child, sister and mother elements as far as they have no id references excluding them from the local context.

The ODD document

The most important standard for encoding texts in the humanities today is the TEI guidelines. In the on-going work towards a version 5 of these guidelines, the TEI council is re-designing its authoring tool for the standard, introducing the ODD system. This system makes it easy to write the documentation, the formal description and examples into the same source document, from which DTDs or XML schemas can be automatically extracted, as well as documentation in HTML and PDF.

The ODD document is a TEI document. It is used to write the documentation for an encoding schema. In this documentation, certain special elements are included. These elements are used by a software package called Roma to extract a DTD, an XML schema and a RLN schema based on the ODD document.

To show how this is done, see this sample from our ODD:

```
<div2>
 <head>Museum numbers</head>
  <p>Elements to encode the museum numbers.</p>
   <schemaSpec>
[...]
    <elementSpec
     xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
     xmlns="http://www.tei-c.org/ns/1.0"
     xml:id="mnr" usage="req" ident="mnr"><equiv/>
    <gloss></gloss>
    <desc>To encode one museum number. Replaces the
        museumsnr element.</desc>
    <classes>
   <memberOf key="se_avsn"/>
   <memberOf key="se_mnr"/>
    </classes>
    <content>
   <rng:text/>
    </content>
   </elementSpec>
  </schemaSpec>
</div2>
```

When this section is read by the Roma software, documentation is created in HTML and PDF based on the textual description, also including documentation of the formal structure defined in the ODD. In addition, a line defining the **mnr** element is written to the DTD, with a content model consisting of text and with the attributes included in the attribute classes **ae_avsn** and **se_mnr**.

Using the ODD system, we are also in a position to include the acquisition catalogue in a TEI document. This would be a TEI quite different from most other TEI documents, but it will still be usable for users knowing TEI and for software tailor-suited for TEI documents. Some users may just exclude the special elements defined for our use and use the catalogues as typical TEI documents.

To include our data description in a TEI data description, the following steps will have to be taken:

1. Define the root element of our special encoding scheme, the **textPart** element, as a text level element in TEI that can be used wherever a paragraph element (**p**) can be used.
2. Include our ODD as a module in a TEI data definition and run it through the Roma software.

This will result in a valid TEI data description and documentation, in which our special element **textPart**, with all its children, can be used in the same way as a **p** element is used in other TEIs.

Extracting a CIDOC-CRM compatible model

Readers familiar with the CIDOC-CRM ontology (CIDOC, 2003) will already have seen how the elements defined above fit into the ontology. This is nothing more than should be expected when a formalised reading of museum information from a core museum document class is described.

In the further use of the XML documents encoded in relation to the data model, the information from each **textPart** is used to build a main artefact object in the database, with the elements of the sub-tree stored in various tabled in the relational database as described in the model and the mapping from the XML model to the database model (Jordal forthcoming). By keeping the IDs from the XML document in the database, a link back to the source is kept.

This is important to facilitate examination of data quality later on. But it is also helpful if the TEI version of an acquisition catalogue is linked together with other TEI documents. In that case, the storing of the IDs from the XML document in the database makes it possible to track the artefact information from the TEI document into the database.

While this do not seem to be especially important now, it will be more and more valuable as the data in the database changes. So, by starting from e.g. a digital version of a letter the reader can follow a link to an acquisition event in the TEI version of the acquisition catalogue. Based on the ID link, the artefacts described can then be followed into the museums catalogue in the relational database, and the current location of the artefacts can be examined, together with conservation and exhibition history.

Conclusion

By using flexible data descriptions and a combination of computer-based processes and human labour, we have been able to SGML and XML encode large acquisition catalogue series. In using the ODD system for our data descriptions, we have gained in three ways. First, we have been able to write better technical documentation and data descriptions because of the help given us by the Roma software. Secondly, we can easily use the TEI standard to encapsulate our acquisition catalogue documents, thus storing them in a standardised way, which is important both for long-time preservation and for use by others. It also gives us an easier way to interconnect our various documents, as many kinds of texts ranging from historical documents, literary works and corpus texts to dictionary editions held by our unit is more and more stored as TEI documents. The third gain by this method is that the ODD document gives a better overview of the structure of the document type than a DTD does. The tools developed in the TEI community helps us create both machine readable and human readable output from the single ODD document. This simplifies the

process of understanding and formally expressing the structure of our documents.

References

CIDOC (2003) Definition of the CIDOC Conceptual Reference Model / Produced by the ICOM/CIDOC Documentation Standards Group, continued by the CIDOC CRM Special Interest Group. ISO/DIS 21127. URL: http://cidoc.ics.forth.gr/definition_cidoc.html (as of 2006-04-05)

Holmen, J. et.al. (1996) "Getting the most out of it - SGML-encoding of archaeological texts." Paper at the IAAC'96 Iasi, Romania. URL: http://www.dokpro.uio.no/engelsk/text/getting_most_out_of_it.html (as of 2005-11-14).

Holmen, J. et.al. (forthcoming) "From XML encoded text to objects and events in a CRM compatible database. A case study". In: *Beyond the Artifact. Proceedings of CAA 2004, Computer Applications and Quantitative Methods in Archaeology.*

Jordal, E. et.al. (forthcoming) "From XML-tagged Acquisition Catalogues to an Event-based Relational Database". Proceedings fra CAA 2004, Computer Applications and Quantitative Methods in Archaeology.

TEI P5 (2005) *Guidelines for Electronic Text Encoding and Interchange. [draft] Version 0.2.1.* TEI Consortium, 2005.

TEI web URL: http://www.tei-c.org/ (as of 2006-04-30).